Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation

Daphne Theijssen, Louis ten Bosch, Lou Boves, Bert Cranen and Hans van Halteren

Abstract

In existing research on syntactic alternations such as the dative alternation, (*give her the apple* vs. *give the apple to her*), the linguistic data is often analysed with the help of logistic regression models. In this article, we evaluate the use of logistic regression for this type of research, and present two different approaches: Bayesian Networks and Memory-based learning. For the Bayesian Network, we use the higher-level semantic features suggested in the literature, while we limit ourselves to lexical items in the memory-based approach. We evaluate the suitability of the three approaches by applying them to a large data set (>11,000 instances) extracted from the British National Corpus, and comparing their quality in terms of classification accuracy, their interpretability in the context of linguistic research, and their actual classification of individual cases. Our main finding is that the classifications are very similar across the three approaches, also when employing lexical items instead of the higher-level features, because most of the alternation is determined by the verb and the length of the two objects (here: *her* and *the apple*).

1. Introduction

Choice is present in language production in many forms, for instance in the choice of words, intonation contours and sentence structure. A common example of the latter is the dative alternation in English, in which speakers and writers can choose between a prepositional dative construction (example 1) and a double object construction (example 2). Both constructions contain two objects, being the 'theme' (*the poisonous apple*) and the 'recipient' (*Snow White*).

(1) The evil queen gave the poisonous apple to Snow White.
(2) The evil queen gave Snow White the poisonous apple.

There is already a vast body of research on the English dative alternation. For the last decade or so, researchers have used different types of features suggested in the literature, and combined them in multivariate models to predict the construction used in corpus data (e.g. Arnold et al. 2000; Gries 2003; Bresnan et al. 2007). The predictive features used were higher-level features that were syntactic, semantic and discourse-related in nature (as opposed to lexical features such as the actual words used). The features mostly represent characteristics of the recipient and the theme, e.g. indicating whether they are animate or inanimate, or whether they are previously mentioned in the discourse (*discourse given*) or represent new information (*discourse new*).

Such higher-level features are commonly used in research on syntax because they help to find general patterns in syntax, beyond the individual words and their frequencies. Another reason for the common use of these features in syntactic research is their predictive power. Bresnan et al. (2007), for instance, reached an accuracy above 90% when testing on previously unseen data, compared to a majority baseline of 78%.

The research presented in this article starts from Bresnan et al. (2007), and also takes a corpus-based approach to studying the dative alternation, making use of the same predictive features suggested in the literature. The success of Bresnan et al.'s regression model with higher-level features has inspired many other researchers investigating syntactic alternations to use the same approach. It has proven to be a very useful approach to find patterns in corpus data and experimental data, which has led to interesting insights in the syntactic choices people make, across different genres (Hinrichs and Szmrecsányi 2007; Szmrecsányi and Hinrichs 2008; Tagliamonte and Jarmasz 2008; Jankowski 2009; Szmrecsányi 2010), and varieties of the same language (Mukherjee and Hoffmann 2006; Bresnan and Hay 2008; Grimm and Bresnan 2009; Bresnan and Ford 2010; Kendall et al. 2011; Wolk et al. 2012). It appears that, all else being equal, people tend to place animate noun phrases before inanimate ones, definite before indefinite, discourse given before discourse new, pronominal (having a pronoun as its head) before nonpronominal, first and second person (*me*, *you*) before third person, and shorter before longer.

In linguistic research, as exemplified by the studies mentioned above, the goal is to find a model that describes language data accurately, and that tells us something about the roles that certain linguistic features play. The modelling technique commonly used in previous linguistic research,

logistic regression, is attractive for several reasons. First of all, it is a multivariate approach: it enables us to investigate the contribution and significance of several features at the same time. Second, contrary to alternative classifiers such as LDA (e.g. Gries 2003) that make strong assumptions about the statistical distributions of the data, regression models are able to deal with non-numerical data. This is beneficial since nominal (often binary) data is very common in corpus studies on syntax. Third, the models themselves are fairly simple; they provide coefficients that indicate the relative roles that the individual features (values) play. Fourth, multiple regression models make it possible to combine fixed variables (the features) and random variables (random effects). This combination helps to establish the effect of the linguistic variables of interest, while controlling for random variables that are usually not of primary interest, such as the individual speaker.

However, there are also some problems with these regression models. One of the major drawbacks of logistic regression is that it requires certain properties of the data that cannot always be fulfilled. Features should be independent, for instance, but in reality they are are often correlated. For example, it is known that the dative alternation is influenced greatly by the relative lengths of the recipient and the theme, but also that humans tend to place pronouns *before* full noun phrases in the clause. These two features, length and pronominality, are correlated because pronominal objects are usually short, i.e. consisting of a pronoun only. Correlated features cause problems with the interpretation of the roles that the individual features play in the model. For example, correlations can cause coefficients to flip sign or loose statistical significance. This means that the effect of pronominality in the model could become insignificant or receive a coefficient that indicates the opposite of the direction expected (on the basis of existing research), because most of its variance is already explained by the length feature. Such correlation issues obviously increase the risk of misinterpreting the effects. There are many mathematical approaches to solve the problems caused by collinearity, for example by centering or residualising variables (for details, see for instance Baayen 2008). Such approaches mostly involve some form of transformation of the original data into data that has the required characteristics. However, if length difference is for instance residualised on the pronominality of the recipient and the theme,[1] the feature under investigation is not length difference itself, but this less straightforward residualised version. Linguists often want to answer research questions about certain features, and transformations tend to hamper the interpretability of the models in

terms of the original data. What we thus need are different modelling approaches that do not suffer from these problems.

There is another reason why we want to move beyond regression models. Syntacticians with various backgrounds are now taking more and more interest in the social and cognitive aspects of language. There are, for instance, recent multivariate approaches that combine the results of sociolinguistic studies, researching the effect that factors such as age and gender may have on language, with cognitive linguistic studies (e.g. Geeraerts et al. 2010). Also there are attempts to relate findings from corpus studies to observations in psycholinguistic experiments (e.g. Bresnan and Ford 2010). Multivariate models such as regression models can successfully be exploited for the purpose of analysing the relative importance of higher-level features in specific data sets under investigation, but these models cannot elucidate the role of these features in cognitive processes. The first goal of the present article is therefore to investigate the explanatory power of the higher-level features in a model that is more likely to be cognitively plausible than regression models. Several approaches have been developed for this purpose, of which connectionist models are perhaps the best-known (e.g. McClelland et al. 2010). Although connectionist models have gained substantial interest in psycholinguistics, they have less traction in formal linguistics, probably because the internal structure of these models is opaque. Recently, Baayen (2011) used Naive Discriminative Learning to model the dative alternation with higher-level features.

In this article we use yet a different approach: *Bayesian Networks* (Pearl 1988). This approach is fully transparent and does not make assumptions about the statistical distributions of the predictor variables. Bayesian Networks make it possible to integrate possibly uncertain prior knowledge and possibly erroneous empirical evidence of different types and different sources in a consistent probabilistic framework (cf. Section 3.2 for more detail). Integrating partial and noisy sensory input and volatile procedural and semantic memory is what the brain does all the time, especially in the initial stages of the processing where not all information is available yet. Therefore, Bayesian Networks form an attractive analogue for cognitive processes (Chater et al. 2006, 2010).

Computational grammar learning models using Bayesian inference have already been shown to be able to learn the dative alternation in a small set of relatively simple, artificial sentences, making use of grammar rewrite rules only, without any higher-level information (Dowman 2004). The question therefore arises whether the higher-level features are really necessary for explaining the dative alternation and for generalising from

(small) data sets to actual language production. This brings us to our second research goal: to investigate the suitability of a model that can claim cognitive plausibility and that is not provided with higher-level features, but with lexical items. To that end, we adopt a *memory-based learning* approach (Daelemans and van den Bosch 2005), in which learning is defined as the storage of some sort of representation of experience (cf. Section 3.3 for more detail). This memory of previous experience is then used to guide actions in new situations. For the dative alternation, this means that humans learn the contextual suitability of the two constructions by storing some representation of the occurrences they produce themselves and hear or read in other people's language use. In the context of current discussions about the existence of an innate, specifically language-related ability, it is interesting to note that memory-based learning has no need to assume an innate language faculty. Language, according to this theory, is learned from input only, making use of the general cognitive abilities that we possess. The underlying idea of this model therefore shows many similarities with exemplar-based models of language processing (Gahl and Yu 2006), and with for instance data-oriented parsing approaches (Bod 2009). When storing all experience with the dative alternation, there is no reason to abstract away from the original input that we hear by defining higher-level features. This makes the role of the higher-level features used in existing research unnecessary and, using Occam's razor, implausible. The only assumption we need to make for studying the dative alternation in the way we do, is that humans have learned the meaning of a number of verbs and the existence of the semantic roles 'recipient' and 'theme'. Memory-based learning does not make assumptions about statistical distributions of the items that are kept in memory.

In order to address our two research goals, we will employ two approaches to model the dative alternation that can be associated with cognitive processes: Bayesian Networks and memory-based learning. For the sake of comparability, we also include the traditional logistic regression models. We evaluate the suitability of the three approaches for studying the dative alternation, on the basis of the following three criteria:

- the quality of the model in terms of classification accuracy
- the interpretability of the model in linguistic research
- the actual classification of individual cases by the model

The remainder of this article is structured as follows: In Section 2, we describe the data set and the various features used. The modelling techniques are introduced in Section 3 and they are evaluated according to

the criteria in Section 4. The article ends with our general discussion and conclusion, provided in Section 5.

2. Data

2.1. Data collection

The data set was extracted from the 100-million-word British National Corpus (BNC Consortium 2007), following the semi-automatic approach in Theijssen et al. (2011a), as summarised below. For more details, refer to Theijssen et al. (2011a).

We used a Perl script to extract all sentences with an occurrence of a *dative verb*. A list of dative verbs was established in the following way (Theijssen et al. 2011a): 1) extracting 264 dative verbs from various linguistic resources, including VerbNet (Kipper et al. 2000) and the verb classification by Levin (1993), 2) removing the 86 verbs with a frequency below 1,000 in the BNC (e.g. *fax)*, and 3) manually filtering out the 102 verbs that alternate with a preposition other than *to* (e.g. *cook for*) and/or that allow only one of the two constructions (e.g. *inform*). The procedure resulted in a list of 76 dative verbs. All sentences with a dative verb tagged as a verb in the corpus (and not as a noun, as is for instance possible for *offer*) were then parsed with the Functional Dependency Grammar (FDG) parser, version 3.9, developed at Connexor (Tapanainen and Jarvinen 1997). A second Perl script was used to extract all dative constructions from the syntactic parses (152,008 in total), after which we employed two automatic filtering steps.

In the first filtering, we used another Perl script to automatically filter out the 44,464 candidates that had at least one of the following features: 1) the theme or recipient was a clause, 2) the clause was in passive voice, 3) the verb was imperative, 4) the theme or recipient preceded the verb, 5) the verb was phrasal (e.g. *I'll send you out that*), 6) the clause was interrogative, 7) recipient and theme were reversed with respect to the expected order (e.g. *I give to him a letter*), 8) the theme was an adjective, 9) the theme or recipient was empty, 10) the clause was a fixed expression (e.g. *I'll tell you what*), 11) there was more than one verb, theme or recipient (e.g. *I gave it to her and to him*). Most of these filters were used to prevent the influence of other types of syntactic variation than those of interest in this research (passive versus active voice, declarative versus interrogative mode, the placement of adverbials, etc.). Some were used to make sure that the features we want to apply later were applicable (e.g. it is

not possible to establish the definiteness of the theme if it is a clause, not a noun phrase).

In the second filtering, a Perl script was used to remove candidates that were likely to contain parse errors (21,965 in total). We removed the candidates where the recipient or theme lacked the presence of a pronoun or noun. For the double object constructions, we filtered out all candidates in which 1) the last word of the recipient and the first word of the theme were proper nouns (e.g. *give John Smith*), 2) the last word of the recipient was a possessive (e.g. *give Mary's money*), 3) the last word of the theme was a reflexive pronoun (*give it yourself*), 4) the verb was *make*, and both recipient and theme were persons in WordNet (Fellbaum 1998) (e.g. *make him king*), 5) the verb was *take*, and the theme was a time noun in WordNet (e.g. *takes me an hour*), and 6) the recipient and theme together were likely to be one phrase (e.g. *write the professional letters*), based on their co-occurrence in the BNC. For the prepositional dative construction, we excluded all instances where the recipient was a location in WordNet (e.g. *bring him to school*), and where the prepositional phrase was likely to be the complement of the theme rather than the verb (e.g. *give access to the garden*), again based on the co-occurrences in the BNC.

After the filtering, 85,579 dative candidates remained. We next checked over 17,000 candidates manually, removing candidates that contained parse errors and that were not dative constructions. The checked subset contained all candidates from the spoken part of the BNC (>11,000 candidates), and yielded 7,757 confirmed dative constructions. To increase the diversity in the data, we supplemented the spoken material with a random selection from the written material (>6,000 candidates). The resulting data set contains 11,784 instances, of which 7,757 are spoken and 4,027 written, spread over various genres, e.g. public meetings, private conversations, news paper articles and fiction texts.

2.2. Medium and length difference

There are two *basic* features that will be used in all models, both in the models that use the higher-level features and the model that uses only lexical items: (1) Medium (spoken or written) and (2) the length difference between the theme and the recipient. Medium is a binary feature that is easy to establish on the basis of the metadata provided in the BNC. Length difference is used as an approximation of syntactic weight, which is known to play a role because of the principle of end weight (Behaghel 1909).

There are many (often correlated) alternatives for establishing the syntactic weight (Shih and Grafmiller 2011), but in this article we limit ourselves to a number of variations of the length difference in words. Since the Bayesian Network tool we employ is not able to deal with interval data, we also include several ways of discretisation, leading to a total of six definitions:

- LenDif: theme length in words minus recipient length in words
- lnLenDif: the log of the ratio between these two lengths
- dLenDif5: an intuition-based discretised version of LenDif with 5 levels (i.e. similar lengths, a longer recipient, a longer theme, a *much* longer recipient and a *much* longer theme)
- dLenDif6: a frequency-based discretised version with 6 levels
- dLenDif10: a frequency-based discretised version with 10 levels
- dLenDif78: a frequency-based discretised version with 78 levels

The cut-off points for the intuition-based discretisation were chosen so that if the ratio between the number of words in the two objects was ≥1:3, the longest of the two was considered *longer*, and when the ratio was ≥1:4, it was considered *much* longer. The frequency-based discretisation in resp. 6 and 10 levels was based on the frequency distributions of LenDif in the data set. For the 6-level discretisation, each level had a frequency of at least 1,100 instances, and for the 10-level distretisation, each at least 400 instances. In the 78-level discretisation, each level contained one unique value of LenDif, with the number of instances per level varying from 1 to 3,522.

2.3. Verb

It is known that many verbs have a strong preference for one of the two constructions (e.g. Gries and Stefanowitsch 2004). For this reason, all models take into account the verb used in the dative construction. In the memory-based model, the verb is included in the lexical items, as will become clear in Section 2.5. For the regression model and the Bayesian network, the treatment of the verb is explained in Section 3. The 46 verbs used in our research are shown in Table 1,[2] together with their frequencies in the data set we use.

Table 1. Verbs and their frequencies in the data set

give	6974	leave	124	pass	77	throw	48	permit	31
tell	799	lend	120	charge	74	bear	44	deal	30
send	363	cause	113	promise	74	issue	43	advance	23
show	342	write	112	wish	64	award	42	read	19
pay	232	teach	111	grant	62	play	40	vote	13
offer	206	make	98	feed	61	pose	38	forbid	10
do	205	present	92	deliver	59	serve	38		
bring	179	take	91	allocate	54	refuse	36		
sell	158	deny	89	assign	49	accord	35		
owe	152	hand	81	guarantee	48	bid	31		

2.4. Higher-level feature extraction

Two of the three modelling techniques employed in this article make use of the higher-level features suggested in the literature (the third technique, memory-based learning, uses lexical items only). These higher-level features are often difficult to define and to annotate with high agreement levels between human annotators. We solve this problem by making use of automatic feature extraction, so that that the definitions are clear and the annotations themselves consistent (Theijssen et al. 2011a). Moreover, Theijssen et al. (2011b) show that the quality (prediction accuracy) of logistic regression models applied to data annotated with this automatic method is equally good as the models found for data with manual annotations, as long as there are enough data points. Since the data set used in the present article is larger (over 11,000 instances) than the largest set (approx. 8,000 instances) included in Theijssen et al. (2011b), we believe the automatic feature extraction approach to be suitable for the present research. It is explained in more detail below.

All instances in the data set were annotated automatically for eight higher-level features, using the feature extraction Perl script in Theijssen et al. (2011a). The names, definitions and values of the features are summarised in Table 2. All higher-level features are binary.

Table 2. Higher-level features for which the instances have been annotated automatically

Name	Definition	Values (binary)
AnRec	animacy of recipient	animate, inanimate
DefRec	definiteness of recipient	definite, indefinite
DefTh	definiteness of theme	definite, indefinite
GivRec	discourse givenness of recipient	given, nongiven
GivTh	discourse givenness of theme	given, nongiven
PrsRec	person of recipient	1st/2nd (local), 3rd person (nonlocal)
PrnRec	pronominality of recipient	pronominal, nonpronominal
PrnTh	pronominality of theme	pronominal, nonpronominal

For establishing the definiteness of recipient and theme, we used the POS tags available in the BNC. When the head (as found in the syntactic parse) occurred with a definite article, it was classified it as *definite*. The same applied to a head that was, or occurred with, a demonstrative, interrogative, relative or possessive pronoun. Similarly, the script considered *definite* heads that were a reciprocal, reflexive or personal pronoun, or a proper noun.

With respect to the discourse givenness of the theme and the recipient, the approach taken is as follows. Given the fact that indefinite objects are mostly new to the discourse, the script classified all indefinite objects as *discourse new*. Definite objects of which the head was a personal pronoun, and of which the head was preceded by a demonstrative pronoun, were labelled *discourse given*. For the remaining definite objects, the script checked the preceding contexts, with a maximum length of 20 clauses (i.e. until the 20th preceding word that was tagged as main verb). If the head itself, or a synonym of the head was found within this preceding context, the object was considered *discourse given*. We used the synsets in WordNet to extract the synonyms. The remaining definite objects were given the value *discourse new*.

For the person of the recipient, the script simply checked whether the syntactic head is a first or second person pronoun.[3] If this was the case, the recipient was *local*, otherwise it was *non-local*.

For the pronominality of the recipient and the theme, the script checked if the head (as found in the syntactic parse) had a POS tag for (any type of) pronoun. If so, the object was classified as *pronominal*, and if not, as *non-pronominal*.

2.5. Extracting lexical items

For the memory-based approach, we use two different variants of lexical items as features, being word forms and lemmas. Since the FDG parser provides lemmas in its output, we extracted the lemmas directly from the FDG parses. As already mentioned in Section 1, we assume that humans know the meaning of a number of verbs and the semantic roles 'recipient' and 'theme'. As features, we therefore use specific lexical items present in the recipient, the theme and the verb. Consider sentences 3 and 4:

(3) I gave a dog biscuit to it.
(4) I gave it a dog biscuit.

The word forms extracted from these sentences would be:

- the verb: *V:gave*
- the recipient head: *Rh:it*
- the beginning of the recipient: *Rb:it*
- the theme head: *Th:biscuit*
- the beginning of the theme: *Tb:a*

For the recipient and the theme, we used a Perl script to extract the head from the dependency parses, as well as the first word or lemma (after removing the preposition *to* in prepositional dative cases). The reason for including the beginning of the recipient and theme is that previous research has indicated that definiteness seems to play a role in the dative alternation. Since definiteness is mostly determined by the presence or absence of certain determiners at the beginning of the object, it may well be that it is not the higher-level feature itself that influences the choice for either syntactic construction, but the presence of certain words or lemmas. We therefore include these lexical items in this model, to see what role they play in the memory-based learning model.

3. Modelling techniques

In this section, we elaborate on the three modelling techniques we use: logistic regression, Bayesian Networks and memory-based learning.

3.1. Logistic regression

In this approach, we employ the eight higher-level features in Table 2 and the length difference, and include them as predictors in a mixed-effect logistic regression model. The Medium (spoken or written) is also included as a predictor, and so are all its interactions with the nine other predictors. The verb of the construction (e.g. *give*) is included as a random factor.

Using the values of the predictors and the verb *i* we establish a regression function that predicts the natural logarithm (ln) of the odds that the construction *C* in instance *j* is a prepositional dative. The regression function is:

$$\ln(\text{odds}(C_{ij}=1)) = \alpha + \Sigma(\beta_k V_{kj}) + e_{ij} + r_i .$$

The α is the intercept of the function. The terms $\beta_k V_{kj}$ contain the weights β and values V_j of the predictors *k*. The random effect r_i established for the verbs (*i*) is normally distributed with mean zero ($r_i \sim N(0, \sigma_r^2)$), independent of the normally distributed error term e_{ij} (e_{ij} $N(0, \sigma_e^2)$). The optimal values for the function parameters α, β_k, r_i and e_{ij} are found with the help of Maximum Likelihood Estimation.[4]

The variable selection method is as follows: We start out with a model including all predictors and two-way interactions with Medium, and remove all insignificant *interactions* in one single step. This step is carried out on the full set of 11,784 instances available, after which only significant predictors remain. We perform this six times, each with a different representation of the length difference. The discrete representations of length difference are interpreted as binary features: one binary feature for each discrete level (except one that is included in the intercept). The representation with 78 levels (dLenDif78), and hence 77 binary features, runs into sparseness problems,[5] but the other five all score *model fit* accuracies that do not differ significantly from each other, training and testing on the same 11,784 instances: 92.2% to 92.5%.[6] Since the log of the ratio between the two lengths (lnLenDif) scores the highest model fit (92.5%), and is one of the two most parsimonious definitions with respect to number of regression coefficients (only 1, because it is numerical), it is selected for further analysis.

3.2. Bayesian Network

The higher-level features that are implicated in selecting a dative construction can be considered as just as many modules in a very complex system that generates sentences. To avoid making things overly complex, we assume that the structure of that system can be represented in the form of an acyclic directed graph, which means that (parent) module M_x can affect the operation of (daughter) module M_y, but not the other way round. Obviously, the fact that we know the direction of the dependencies implies that we claim to have prior knowledge about the structure of the process that we are investigating. It also means that we can draw a picture of the structure of the system in which the modules are represented by nodes, and the connections between the modules are represented by single-headed arrows (cf. Figure 1). An arrow from parent node N_x to daughter node N_y implies that the latter is conditionally dependent on the former: The value of N_x influences the operation of N_y. The beauty of the theory of Bayesian Networks (Pearl 1988) is twofold. First, this theory allows us to learn the strength of the connections between modules from data, which is tantamount to integrating knowledge (represented by the nodes and connections in the network) and empirical observations. Second, the theory allows for efficient computation, because it follows from the structure of the network which modules are conditionally independent. In technical terms: it allows us to factor the joint probability $p(N_1, \ldots, N_m)$ of observing specific values for all m modules (represented as nodes) in the network at a given time in such a manner that only conditional dependencies (represented by arrows in the network) need to be taken into account.

Compared to logistic regression, Bayesian Networks have several advantages. One, which is not relevant in the case of dative alternation, is that the output node (the black node labelled 'Cons' in Figure 1) can take an arbitrary number of values. Second, the structure of the Bayesian Network represents a decomposition of the original modelling problem into smaller subproblems. Third, the way in which the features interact is easier to visualise. There are several public-domain software packages that allow one to easily create and manipulate networks and to visualise the strength of the connections that were learned from training data (e.g., in the form of the thickness of the arrows, cf. Figure 2). Unsurprisingly, these advantages come at a cost: There is no proven method for learning the *structure (topology)* of a network from training data. Incomplete prior knowledge may cause mistakes in drawing the connection scheme and thus result in misleading accounts of the structure of the process that generates the output

observations. For this reason we will explain the decisions that were made in creating the network in Figure 1 in substantial detail.

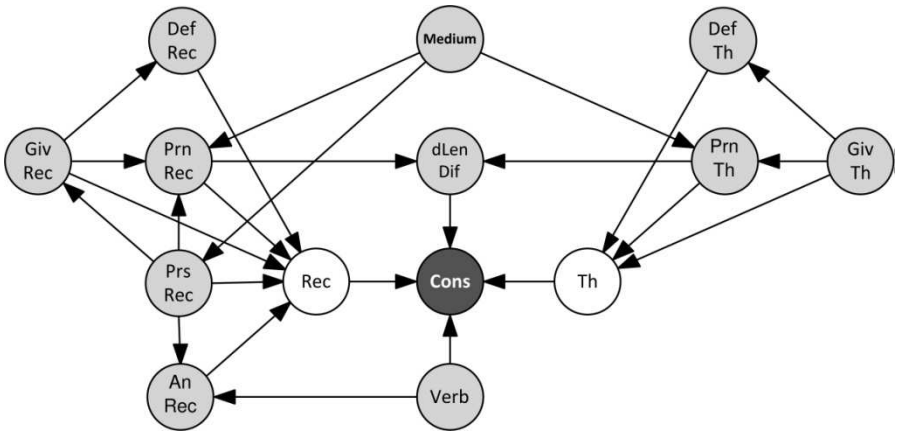

Figure 1. Theoretically motivated Bayesian Network. The grey nodes are observable input nodes, the white ones are hidden nodes, the black node is the (observable) output node.

The names of the nodes are the features in Table 2, supplemented with a node for Length difference (dLenDif) and a node for Verb, which is treated as a discrete variable with 46 (nominal) values. Since the syntactic construction is queried using Bayesian inference on the evidence set for the other nodes, it is indicated with the black *output* node labelled *Cons*. The network contains two hidden nodes in white: *Rec* (for *Recipient*) and *Th* (for *Theme*). These are nodes that have no values in our training or test data, but are nodes that allow the network to combine and 'summarise' the information about the theme and the recipient in a number of states. This is an elegant way to combine the various, possibly correlated, characteristics of the theme and the recipient (the grey *input* nodes), and see the relative influence they have on the recipient and the theme (i.e. the hidden nodes) separately.

Each of the arrows in the network is motivated below (sorted alphabetically by node name). It goes without saying that many arrows could be added and removed, either randomly or on the basis of other linguistic intuitions or theories. However, our goal is not to perform data exploration and find the single best model, but to apply our hypotheses about the dependencies between the features and the syntactic construction used, preferring a network structure that is interpretable and transparent.

3.2.1. Animacy (AnRec)

The animacy of the recipient has no direct influence on the other nodes, as far as we know. Therefore, there is only an arrow from AnRec to the hidden node Rec.

3.2.2. Definiteness (DefRec, DefTh)

As far as we are aware, the definiteness has no direct influence on the other feature nodes, so it only has an arrow towards the hidden nodes Rec and Th.

3.2.3. Length difference (dLenDif)

The length difference is known to have a strong influence on the construction used (e.g. Bresnan et al. 2007; Theijssen et al. 2011a), because of the principle of *end weight* (Behaghel 1909). Therefore, we added a direct arrow from dLenDif to Cons.

3.2.4. Discourse givenness (GivRec, GivTh)

When an object has been mentioned previously in the discourse, we expect that the speaker or writer is more likely to use a pronoun (e.g., referring to a previously mentioned book with *it*), hence the arrows to PrnTh and PrnRec. When the object represents new information, we assume it is more likely to be realised as an indefinite noun phrase (e.g., *a book* is usually a book that has not been mentioned before). Besides arrows to these two features, we also include arrows to the hidden nodes Rec and Th.

3.2.5. Hidden nodes (Rec, Th)

The two hidden nodes receive and collect information from the various feature nodes, and both provide information to Cons.

3.2.6. Medium

Biber (1988) has already shown that spontaneous, usually spoken language contains significantly more pronouns and mentions of first and second persons (*me, you*) than more formal and written language. We therefore added arrows from Medium to PrnRec, PrnTh and PrsRec. The bias towards the double object construction is usually stronger in spoken data (86.6% of the spoken instances in our data set were double object) than in written data (64.3% of the written instances in our data). However, we inspected various existing dative data sets (e.g. those in Theijssen 2010; Theijssen et al. 2011a) and discovered that these differences can all be explained by the relative frequencies of the values for the three features PrnRec, PrnTh and PrsRec. For this reason, there is no direct arrow from Medium to Cons.

3.2.7. Pronominality (PrnRec, PrnTh)

As mentioned in Section 1, the length difference between the two objects is greatly influenced by the pronominality of these objects. The reason is that pronominal objects are often very short because they usually consist of a pronoun only. For this reason, the network includes arrows from PrnRec and PrnTh to dLenDif.

3.2.8. Person (PrsRec)

When the recipient is in first or second person (local), it is almost always an animate, pronominal, discourse given recipient (*me, us, you*). Because of this direct influence, the network includes arrows from PrsRec to AnRec, GivRec and PrnRec. There is also an arrow to the hidden node Rec.

3.2.9. Verb

As mentioned previously, many verbs have a strong preference for one of the two constructions (Gries and Stefanowitsch 2004). Therefore, there is a direct arrow from Verb to Cons. Also, some verbs may influence the likelihood of the animacy of the recipient. For instance, in various existing dative data sets (e.g. those in Theijssen 2010; Theijssen et al. 2011a), we saw that the verb *show* is more likely to occur with an animate recipient,

since one usually shows something to people, not to things. This explains our choice to include an arrow from Verb to AnRec.

3.2.10. Methodology

The network was designed in the Windows user interface GeNIe, a modeling environment for graphical decision-theoretic models developed by the Decision Systems Laboratory of the University of Pittsburgh.[7] The parameter learning on the training data and the inference on the test data was performed in GeNIe's underlying reasoning engine SMILE (Structural Modeling, Inference, and Learning Engine). SMILE is a library of C++ classes implementing graphical decision-theoretic methods such as Bayesian networks and influence diagrams.

Since the goal of the present article is to present an overall evaluation of the suitability of Bayesian Networks for modelling the dative alternation, we decided not to perform any tuning of the tool, employing GeNIE/SMILE's default settings instead. By default, the parameter learning is done with Expectation Maximisation with randomised initial parameter settings. For each test case, the evidence of the nodes was set to the feature values in question, after which the beliefs in the network were updated through the default inference approach (the clustering algorithm). The probability assigned to the node *Cons* was then used to classify the case, choosing the class with the highest probability in the histogram provided for the two possible outcomes.

With respect to LenDif, GeNIe/SMILE was able to deal with the discretised versions only, hence the label *dLenDif* (with a *d* for discretised) in Figure 1. For the hidden node *Th*, we tested all seven possible numbers of values, given the binary input from the three parent nodes: 2, 3, 4, 5, 6, 7 and 8. The same numbers were tested for *Rec*, supplemented by 16, 24 and 32 because of the higher number of parent nodes (5) and hence the high number of possible input combinations (2^5=32). To explore the effect of different cardinalities of the nodes Th, Rec and dLenDif, we thus tried 7 (Th) * 10 (Rec) * 4 (dLenDif) = 280 combinations. Note that all models are the same in their network structure; they only differ with respect to the number of values possible for some nodes. To find the optimal settings, we learned and predicted the same data set with all 11,784 instances.[8] There were 159 combinations that yielded prediction accuracies which did not differ significantly from the highest accuracy (95.1%), i.e. they were all within the 95% confidence intervals according to a binomial distribution. Two of these combinations represented the most parsimonious

representation (requiring only 12 values in total): 5 for dLenDif, 4 for Rec and 3 for Th, and 6 for dLenDif, and 3 for both Rec and Th. We only present the results for the former, since it yielded the highest accuracy (94.5%, compared to 94.3% for the latter option).

3.3. Memory-based learning

Logistic Regression and Bayesian Networks have in common that they use the training data to learn generative models that, given the values of a set of parameters of a new observation, can predict the class to which that observation belongs. Learning the models requires substantial effort and expertise, more often than not expertise at a level that cannot reasonably be expected from naive language users. For example, in this article we do not include the feature 'Concreteness of the theme', which does appear in Bresnan et al.'s model (2007), because of the problems we experienced in annotating that feature (see Theijssen et al. 2011a). Generative models also run into trouble if some feature values occur rarely in the training data (cf. section 3.1.) Memory-based learning as defined by Daelemans and van den Bosch (2005) is a machine-learning method that is designed to avoid problems with labelling data on an abstract level, as well as with sparse observations. Memory-based learning does exactly what its name says: Training examples are stored in the form in which they are observed in text or speech. The only mandatory annotation is the label of the class of which the examples are a member. All training examples are characterised by a number of simple, theory-neutral features, such as the identity of words in a phrase, the identity of the left and right neighbour of a word, the number of syllables of a word, etc. When a new observation comes in to be classified, the examples stored in the memory are searched for items that are most similar (in terms of the features) to the new observation. Learning now consists of finding the similarity measure that minimises the classification error for the training data. Because it does not rely on any kind of generative model, memory-based learning can deal with low-frequency events, even if these represent sub-regularities.

For memory-based learning, we included the two types of lexical items described previously, together with the Medium and one of the six versions of length difference (each version tested in a separate model). The implementation we employed is the nearest neighbour (kNN) classifier in TiMBL (Daelemans et al. 2010). TiMBL stores classified (training) data, and the items in the test set are assigned the class of the nearest neighbour in the stored data. We used the leave-one-out setting, which is a procedure

of iteratively training on all-but-one instances, and testing on the one remaining instance.

TiMBL can be tuned by setting a number of hyperparameters, including the distance metric used for each feature (*m*), the feature-weighting method (*w*), the number of nearest neighbours used for extrapolation (*k*) and the type of class voting weights that are used for extrapolation from the nearest neighbor set (*d*). For each of the twelve lexical item/LenDif variants, we separately tuned these hyperparameters with the help of the wrapper Paramsearch (van den Bosch 2004). Paramsearch finds the best settings by cleverly trying out parameter combinations on subsets of the data. We provided Paramsearch with all data instances and saved the settings that were chosen as 'optimal'. These settings were next used in the leave-one-out setting mentioned above: *m* = Jeffrey divergence, *w* = Gain Ratio, *k* = 9 and *d* = normal majority voting (i.e. all neighbours have equal weight).

All combinations of the type of lexical item (lemma or word) and the definition of length difference yielded an accuracy between 92.4% and 93.1% when training and testing in leave-one-out mode. Since the lemma-based features are more parsimonious (5,563 different lexical items) than the word-based features (6,358 different lexical items), we focus on the lemma-based models. From these, we have selected the model that yielded the highest accuracy (93.1%) for further analysis, which was the model using the discretised version of length difference with 10 levels (dLenDif10).

4. Evaluating the approaches

4.1. Quality of the model in terms of classification accuracy

We evaluate and compare the predictions made by the various models by using the models as classifiers and establishing the percentage of correctly classified instances (the accuracy). We did this in two ways: (1) training and testing on all instances (leave-one-out for Memory-based learning), yielding the model fit, and (2) training and testing in 10-fold cross-validation, using the same division in 10 folds across the approaches. In the 10-fold cross-validation, we re-used the output of the variable selection and hyperparameter tuning applied to all data instances (as described in the previous Section).[9] The model fit accuracies and the average 10-fold accuracies reached can be found in Table 3.

Table 3. Accuracies and their confidence intervals (for model fit) or two times the standard deviations (for 10-fold cross-validation), found for the two baselines and the three modelling approaches

Approach	Features	Model fit	10-Fold cv
Class-majority baseline	none	79.0% (±0.7%)	79.0% (±2.1%)
Verb/LenDif baseline	basic	89.6% (±0.6%)	89.3% (±2.4%)
Logistic regression	basic+higher-level	93.5% (±0.4%)	93.2% (±1.2%)
Bayesian Network	basic+higher-level	94.5% (±0.4%)	93.2% (±1.3%)
Memory-based learning	basic+lexical	93.1% (±0.5%)	92.5% (±1.5%)

For the model fit (leave-one-out for memory-based learning), the three models perform much better than the class-majority baseline of 79.0% (always selecting the double object construction). They are only slightly, but significantly, more accurate than the Verb/LenDif baseline, using the verb and length difference only (89.6%).[10] As mentioned previously, many verbs have a strong preference for one of the two constructions (e.g. Gries and Stefanowitsch 2004). Also, the length difference, which could be interpreted as an approximation of the principle of end weight (Behaghel 1909), is known to have great influence.

When training and testing on all items (the model fit), the best results are reached with the Bayesian Network using the higher-level features (94.5%). In 10-fold cross-validation, the three approaches do not differ significantly, yielding accuracies of 92.5% or higher. The standard deviations for the three approaches are remarkably smaller than those found for the two baselines, which means that the addition of higher-level features or lexical items has led to more stable models. It is interesting to see that a memory-based model, which uses only the basic features and lexical items, is so accurate at predicting the construction used. This is a reason to call into question the importance of higher-level features in language processing. Also, it adds to the questioning of the need for an innate, specifically language-related ability, since memory-based learning explicitly assumes that language is learned from input only, making use of the general cognitive abilities that we possess.

As mentioned in Section 1, the goal in linguistic research is not to find the best performing model, but to find an approach that is sufficiently accurate to constitute a plausible explanation of the underlying cognitive processes and that, at the same time, is able to teach us something about linguistics. The models that we investigate all show a high accuracy; therefore, we keep all three in a more qualitative evaluation.

4.2. Interpretability of the model in linguistic research

In this section, we will evaluate the interpretability of the models in linguistic research, treating them each in a separate subsection.

4.2.1. Logistic regression

The coefficients found for the fixed factors in the logistic regression model are presented in Table 4. What we can learn from the model is that all predictors are significant except Medium, which is kept in because of its significant interaction with DefTh. The fact that so many predictors are significant is not surprising given the large number of data instances. The coefficients in the model can be interpreted because they directly influence the log of the odds that the construction used is the prepositional dative. So, if the recipient is inanimate, the odds increase with 1.03. On the other hand, if the recipient is pronominal, the odds decrease with 1.29, thereby increasing the odds that the construction used is the double object.

Table 4. Coefficients and their properties in the logistic regression model

Feature	Coefficient	Std error	z value	Pr(>|z|)	
(Intercept)	1.14	0.39	2.93	0.003	**
AnRec=in	1.03	0.11	9.37	0.000	***
DefRec=in	0.92	0.14	6.79	0.000	***
DefTh=in	-1.23	0.16	-7.67	0.000	***
GivRec=non	0.86	0.14	6.1	0.000	***
GivTh=non	-1.44	0.15	-9.37	0.000	***
LenDif	-2.3	0.08	-27.16	0.000	***
PrnRec=p	-1.29	0.15	-8.67	0.000	***
PrnTh=p	1.32	0.12	10.78	0.000	***
PrsRec=non	0.33	0.12	2.68	0.007	**
Medium=w	-0.05	0.14	-0.33	0.741	
DefTh=in, Medium=w	0.55	0.17	3.12	0.002	**

Our regression model confirms that animate objects are usually mentioned before inanimate objects, definite before indefinite, discourse given before discourse new, shorter before longer, pronominal before nonpronominal and local (1st/2nd person) before nonlocal (3rd person). As mentioned in Section 1, the fact that regression models are fairly

straightforward is one of the reasons that they have become so popular among syntacticians studying alternations.

It is unclear, however, how the model has dealt with the correlations between the features. The collinearity in the data can be measured with the help of the condition number (*c*-number). For the features in our data, the *c*-number[11] is 14.20, which indicates that there is medium collinearity. In models fitted to smaller data sets, effects of collinearity can become apparent because not all features reach significance. For a large data set such as ours, this is not the case: all features (except Medium) are highly significant. Collinearity can also cause coefficients to flip sign: if two predictors are (strongly) correlated, the predictor with the highest correlation with the criterion will leave only a residual to explain by the predictor with the weaker correlation. The correlation with the residual may have the opposite sign. Seeing that the patterns found are consistent with those found in the vast body of research (including studies using experimental data, and studies investigating the features one at a time), it seems there is no clear influence of collinearity. Still, comparing the actual values of the coefficients, and thereby the relative influence of the feature on the construction used, is not advisable. Another motivation for refraining from a comparison of the coefficient sizes is the fact that most of the statistical variance is explained by the random effect verb and the feature length difference, reaching a model fit accuracy of 89.6%. This means that the coefficients for the other features have only a minor influence on the eventual classification.

4.2.2. Bayesian Network

The Bayesian Network that we used was already presented in Figure 1. In the user interface GeNIe, it is possible to calculate the strength of influence per arrow, and represent this visually in the network. By default, this strength is considered equivalent to the extra information obtained by knowing the value of the parent, compared to the situation where this information is not available. Since in our case each node is characterized by a discrete probability distribution specifying the probability for, say, *N* different values that a node can take, the strength can be represented as the Euclidean distance between the conditional probability distribution of a node given the parent node and the a-priori probability of the node (Koiter 2006):

$$E(node, parent) = \sqrt{\left(\sum_{n=1}^{N} \left(P_n(node|parent) - P_n(node) \right) \right)^2} / \sqrt{2}.$$

where $P_n(.)$ represents the n^{th} component of the discrete probability distribution $P(.)$. Since $P(.)$ is – by definition – a unit length vector, the maximum distance between $P(node|parent$ and $P(node)$ is equal to $\sqrt{2}$ (which is obtained if the two probability vectors are orthogonal, a fact that is easily verified for the two dimensional vectors [1, 0] and [0, 1]). Thus, the division by $\sqrt{2}$ ensures that the resulting distance is between 0 and 1. The strength of influence represents a kind of 'local information gain' yielded by the evidence provided by the parent.

The strengths are shown in Figure 2 by the thickness of the arrows. The figure shows that many of the correlations between the features show thick arrows, indicating they are strongly determined by the value of their parent nodes. This is exactly what we expected. Also, we see that the influence of the features on the two hidden nodes Rec and Th has a very similar strength across these features. It therefore seems that the features are similar in their informativeness for the hidden node, and that the hidden node indeed nicely summarises the information of various correlated features. There are only minor differences in the thickness of the arrows: for both Rec and Th, for instance, the node for givenness (GivRec or GivTh) is one of the more influential.

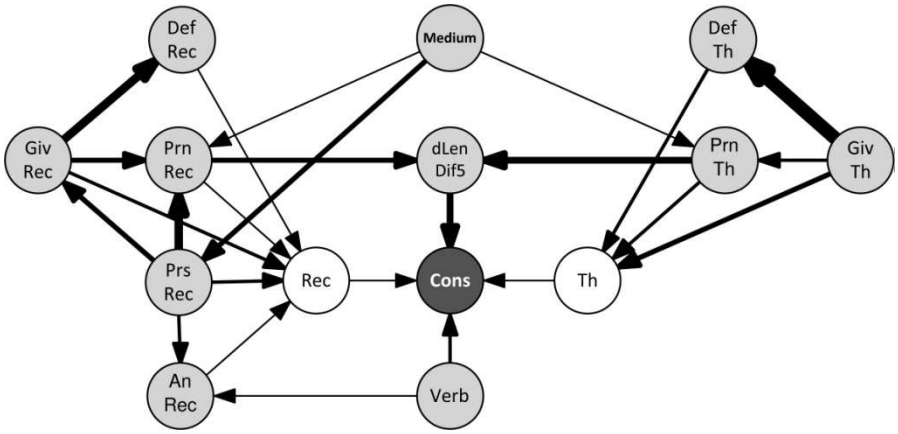

Figure 2. Strength of influence in the Bayesian Network (parameters learned on all data). The thickness of the arrows represents the 'local information gain' of knowing the value of the parent node (for details, see the text).

When we look at the output node Cons, we see that the characteristics of the recipient and the theme have a weaker influence on Cons than the verb

and the length difference. This is as we expected, especially after seeing the high score for the Verb/LenDif baseline in Table 3.

We should note that although the thickness of the arrows nicely visualises which nodes are strongly determined by their parent nodes, the thickness only represents the strength of influence at that (local) place in the network. Consequently, the strengths in Figure 2 do not indicate which arrows are most relevant in the classification task (predicting which Cons was used). For this reason, we established the model fit accuracy, i.e. training and testing on all 11,784 instances, of networks in which we removed one of the arrows. This procedure revealed that the accuracy only dropped significantly when removing one of the four arrows connected directly to Cons. Removing any of the other arrows, even the very thick ones such as that from GivTh to DefTh, did not yield a model fit accuracy that differed significantly from the original network with that arrow. This confirms the general observation that many of the features (here represented as nodes) overlap in their explanatory power: for a large part, they provide the same information. The model fit accuracy decreased most when removing the arrow from dLenDif to Cons (namely to 91.1%), closely followed by the arrow from Verb to Cons (91.5%) and from Th to Cons (91.6%). Removing the arrow from Rec to Cons led to an accuracy of 92.5%. The fact that Verb and dLenDif are very informative is not surprising, seeing our findings in the previous sections.

4.2.3. Memory-based learning

Compared to logistic regression and Bayesian Networks, the memory-based learning model does not allow an easy interpretation at a more general and abstract, linguistically meaningful, level. The only thing we can deduce from the TiMBL output is the Gain Ratio and Information Gain of the *individual* basic and lexical features, as provided in Table 5. The Information Gain measures the difference in uncertainty (i.e. the entropy) between the situation where the feature value is known, and the situation where only the a-priori probability of the class (the dative construction) is known. It is thus very similar to the influence strengths in the Bayesian Network. The Gain Ratio is based on the Information Gain, but normalises it for features with different numbers of values (by dividing the Information Gain by the entropy of the feature values). Only the Gain Ratios are actually used in the model, i.e. as weights in the feature-weighting metric selected in the hyperparameter tuning. The features in Table 5 are therefore sorted according to their Gain Ratios.

Table 5. Individual features, their number of values, Gain Ratios and Information Gain (provided) in the memory-based model (trained on all data).

Feature	Nr of values	Gain Ratio	Information Gain
dLenDif10	10	0.097	0.275
Rh	1,464	0.067	0.350
Rb	888	0.063	0.275
V	46	0.050	0.149
Medium	2	0.050	0.047
Tb	1,032	0.048	0.257
Th	2,133	0.040	0.367

When we look at the Gain Ratios, we see that the length difference receives the highest feature weight. The verb (*V*) ends only in the middle of Table 5 in the ranking for Gain Ratio, and only Medium has a lower Information Gain. This is surprising since we know from previous research (e.g. Gries and Stefanowitsch 2004) and from the Verb/LenDif baselines that the verb is very informative.

The Gain Ratios reveal that especially the characteristics of the recipient weigh heavily in the classification; they are ranked above all other lexical features. So, despite the many possible values for the two features for the recipient (1,464 and 888), knowing the beginning and/or the head lemma is informative. The reason that both recipient features have a high Gain Ratio is probably that for 9,519 instances (80.8%), the recipient consists of one word only, which means that the features *Rh* (head of the recipient) and *Rb* (beginning of the recipient) have the same value: this one word that is the recipient. Of these, 8,465 instances (71.8% of all data) have a recipient that is the personal pronoun *you*, *me*, *them*, *him*, *us*, *her* or *it*. The beginning and the head lemma of the recipient therefore give information about the pronominality (and probably also the short length) of the recipient. The high Gain Ratios thus seem to confirm the finding in previous research that pronominality plays a role in the dative alternation.

The Information Gain values for the two features for the theme are very close to the ones found for the recipient. However, when looking at the Gain Ratio, which takes into account the many values the features can take, we see they are not so informative compared to the other features.

In our description of the lexical items used, we explained that we wanted to include the beginning of the recipient and the theme in order to test whether the relevance of definiteness found in previous research can be explained with the help of lexical items. Table 5 shows that both features representing the beginning of the objects (Rb and Tb) are quite informative

with respect to Information Gain, but only Rb also receives a relatively high feature weight (a Gain Ratio of 0.063, ranked third, compared to a Gain Ratio of 0.048 for Tb, ranked sixth). Based on our observations above, we believe that for the *recipient*, the higher Gain Ratio is most likely caused by the pronominality (and possibly also the length) of the recipient, and not so much by the definiteness. The two most frequent beginning lemmas of the *theme* are the two English articles *a* (3,219 instances, 27.3% of the data) and *the* (1,593 instances, 13.5%). However, since the model output only provides Information Gain and Gain Ratio scores for complete features, and not for the individual feature values that provide information about definiteness, it is not possible to draw any conclusions about the role of definiteness in this memory-based model.

Despite the fact that the memory-based model is difficult to interpret in the sense of understanding which lexical items are most relevant for the choice between the two dative constructions, the model is still useful in the context of linguistic research. Many researchers believe that humans learn language by storing examples, without abstraction in the way it was suggested in traditional linguistic research. Our memory-based model helps to increase the plausibility of this theory.

4.3. Classification of individual cases by the model

Besides evaluating the quality of the models in terms of classification accuracy, and their interpretability in linguistic research, it is interesting to compare the actual classifications made by the models, because they reflect the differences between the models. We do this in two ways: (1) by comparing the classes assigned to the cases, and (2) by comparing the confidence scores provided with these classes.

4.3.1. Comparing the classes

The four panels in Table 6 show four different confusion matrices: one for the 10,837 instances (92.0%) that received the same class from all three approaches, one for the 241 instances (2.0%) for which the class found with Logistic regression differed from the other two, one for the 143 instances (1.2%) for which the Bayesian Network differed, and one for the 562 instances (4.8%) for which Memory-based learning differed. Since the classification problem is binary and we tested only three classification approaches, all data points are covered in the confusion matrices.

Table 6. Confusion matrices of the 11,784 double object (DO) and prepositional dative (PD) instances for which the construction was predicted (Pred=DO or Pred=PD)

a. 10,838 (92.0%) instances classified the same by the three approaches

	Pred=DO		Pred=PD	
DO	8,715	80.4%	116	1.1%
PD	216	2.0%	1,791	16.5%

b. 562 (4.8%) instances classified differently by Memory-based learning

	Pred=DO		Pred=PD	
DO	96	17.1%	162	28.8%
PD	200	35.6%	104	18.5%

c. 241 (2.0%) instances classified differently by Logistic regression

	Pred=DO		Pred=PD	
DO	7	2.9%	146	60.6%
PD	48	19.9%	40	16.6%

d. 143 (1.2%) instances classified differently by Bayesian Network

	Pred=DO		Pred=PD	
DO	36	25.2%	28	19.6%
PD	41	28.7%	38	26.6%

The confusion matrices show that most instances (10,506) receive the same, correct, class in the three approaches: 8,715 double object (DO) cases and 1,791 prepositional dative (PD) cases. So, despite the different modelling techniques of the three approaches, and the different types of features used (lexical and higher-level), the vast majority of the instances is classified correctly in all three approaches. This is not surprising since 89.6% (see Table 3) was classified correctly with Verb and LenDif only, which were both present in the three approaches as well. In fact, of the 10,506 instances that were correctly classified by the three approaches, 94.9% (9,971 instances) was also classified correctly by the Verb/LenDif baseline.

Of the 116 double object (DO) instances that were classified as prepositional dative (PD) constructions by all three approaches, 46 were instances where both the theme and the recipient consisted of a pronoun only (see examples 5, 6 and 7). In total, our data set contained 95 of such double object instances, of which only 12 were correctly predicted by all three approaches. The reason probably is that in examples 5 and 7, the alternative (e.g. *give it to you/him*) is also very common, making it hard to learn when humans use which.

(5) If we give you that we can *give you it* in a certain way, but it is not necessarily meaningful. (BNC: FUL n1285)

(6) but you can always say no to any pack you don't want, you're never under any obligation to buy and we'll stop *sending you them* whenever you ask (BNC: HKD n20)
(7) Well they won't *give him it* straight away, they'll see to you first. (BNC: KCX n1835)

Memory-based learning differs most from the other two approaches (562 instances), and most of these differences lead to misclassification (362 instances). The misclassifications comprise a relatively large proportion of instances containing recipients that are non-pronominal (70.2%), in third person (85.4%), non-given (53.6%) and/or inanimate (35.4%), compared to the rest of the data (24.5%, 51.6%, 20.0% and 14.6%, respectively). Objects in these semantic categories can be instantiated by a much larger number of different words than objects that are pronominal (usually simply one of the pronouns), in first or second person (*me*, *us* and *you*), given (from the limited set of previously mentioned entities) or animate (a person or animal). Since the memory-based learning model makes no use of the higher-level features, but only of dLenDif, Medium and lexical features, it is not very surprising that it performs best at the instances with objects instantiated by more frequent words.

It remains unclear whether the memory-based model fails to classify the more unique instances correctly because of its inability to abstract away from the data (while humans may in fact be doing so), or because its exposure to language data is too small (especially compared to the amount of language to which humans are exposed). Moore (2003) estimated that infants hear approximately 6 million words of speech a year, and adults approximately 14 million. The data we presented to the model was extracted from a corpus of 100 million words. Since we only checked around 20% of the dative candidates found by the parser, the data set could be taken as representative for approximately 20 million words of the corpus. These are words in speech and writing, while the estimate quoted from Moore (2003) was speech only. We can thus safely say that humans, over the years, hear many more dative sentences than the 11,784 we used in the memory-based learning approach.

Logistic regression differs from the other two approaches in 241 cases, most of which are instances where a DO construction is wrongly classified as a PD construction (146 instances). Over 70.5% of these misclassified DO cases were taken from written data, while the percentage of written instances is only 33.7% in the data used in this study. It thus seems that the Logistic regression model is especially tailored towards spoken data (the larger part of the data).

The classification by the Bayesian Network differs least from the other two approaches. The 143 differing cases are spread relatively uniformly in the confusion matrix, showing no clear pattern as to where and why the classification differs.

4.3.2. Comparing the confidence scores

The classifiers not only assign a class label to each case, but also a measure of confidence. In order to compare these measures, we transformed them so that all three represent the likelihood that the construction used was prepositional dative. For Bayesian Networks, we took the probability for the prepositional dative from the histogram provided by GeNIe. For regression, we transformed the log of the odds that the construction was prepositional dative into probabilities. For the memory-based learning models, we used the normalised distributions given in the model output,[12] being values between 0 and 1. The higher this value, the higher the proportion of prepositional datives in the set of nearest neighbours. The three transformed confidence scores will from now on be referred to as *PD-likelihood scores*.

Table 7 presents the pairwise Pearson correlations for the PD-likelihood scores assigned by the three classification models.

Table 7. Pearson correlation between the PD-likelihood scores assigned by the various approaches (for all $p<0.001$)

	Logistic regression	Bayesian Network	Memory-based learning
Logistic regression	1.00	0.95	0.88
Bayesian Network		1.00	0.89
Memory-based learning			1.00

The three correlations are all ≥ 0.88 (indicating high correlation) and highly significant ($p<0.001$). We should note that these high correlations are mostly the result of the fact that the larger part of the data has PD-likelihood scores close to 0 and to 1. The correlations with Memory-based learning are lowest, which shows that the likelihood scores differed most in this approach. There are two possible explanations for this finding: (1) the type of input features used (lexical vs. higher-level) has influenced the PD-likelihood scores, and/or (2) the distribution of the PD-likelihood scores in the memory-based model is different because the scores are proportions,

not probabilities. The proportions differ from probabilities especially because they contain many 0's and 1's, while the probabilities only approximate 0 and 1.

It is to be expected that the PD-likelihood scores assigned to cases that were classified correctly are more at the extremes of the likelihood range (close to 0 and 1), while the scores for cases classified incorrectly are more in the middle (around 0.5). To test if this is true for the three models, we established the average likelihood scores for correctly and incorrectly classified DO and PD constructions. These are presented in Figure 3.

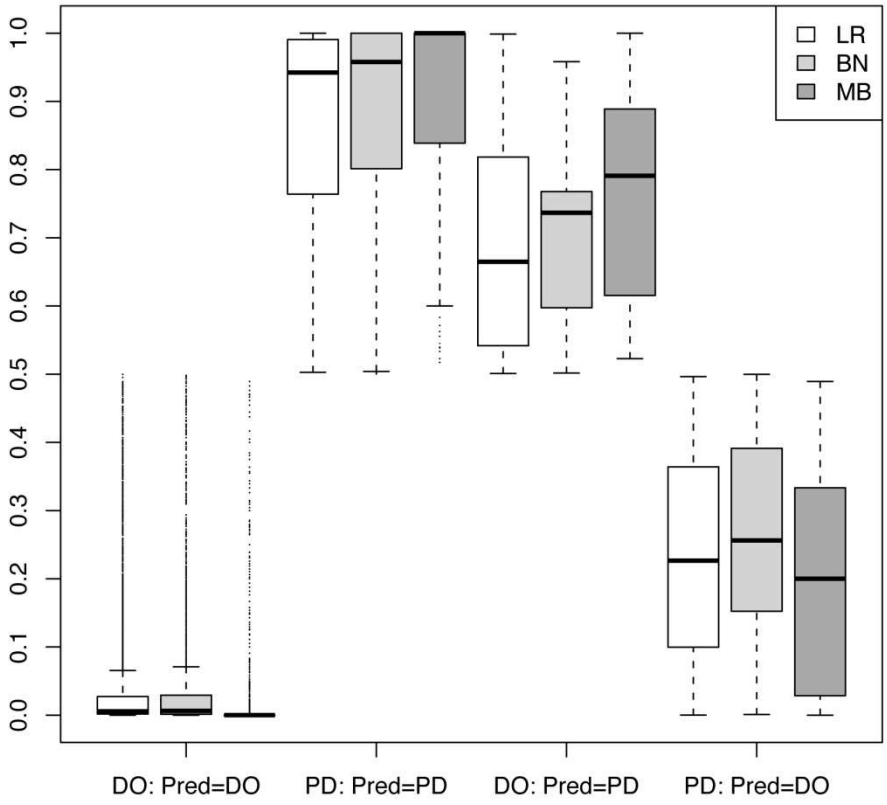

Figure 3. Boxplots of PD-likelihood scores for Logistic regression (LR), Bayesian Network (BN) and Memory-based learning (MB), sorted by the combination of the *actual* dative construction (DO, PD) and the *predicted* construction (Pred=DO, Pred=PD).

The boxplots in Figure 3 show the expected pattern, and are quite similar for the three models. For the correctly classified double object constructions (the bulk of the data), the mean of the PD-likelihood scores is very low, and the quartile boxes very small. This shows that on average, the three models are very certain that the instance is a double object construction. The quartile boxes for the correctly classified prepositional datives, are much broader. This suggests that the confidence of the classifiers is related to the number of positive training examples that are available. Put differently, the confidence for the majority class is higher because the a-priori probability of correct classification is already much higher. At the right hand side of the figure, the two groups of cases that were misclassified receive scores that are approximately equally close to the extremes (0 and 1) as to the middle (0.5). So, despite the fact that the models classified these instances incorrectly, they are fairly certain about the classification, though not as certain as for the correctly classified cases. The likelihood scores are especially extreme for the Memory-based learning model; apparently, in most cases a large proportion of the nearest neighbours represents one of the two dative constructions, which is then selected as the class for the test item.

5. General discussion and conclusion

In this article, we have compared three different approaches to modelling the dative alternation. The first approach was one that is commonly used in linguistics: logistic regression models combining various higher-level features. The second approach used the same features, but a modellling technique that can be associated with cognitive processes: Bayesian Networks. In the third approach, we let go of the higher-level features and employed lexical items in a memory-based learning model.

Logistic regression is a statistical method that is convenient for several reasons: it is a multivariate approach, it is able to deal with non-numerical data, the models are fairly simple and logistic regression makes it possible to combine fixed variables (the features) and random variables (random effects). Also, this article confirms previous findings that logistic regression models with higher-level features are very powerful: 93.2% of the instances were classified correctly in 10-fold cross-validation, compared to a Verb/LenDif baseline of 89.3%. But there are also some drawbacks. First, it is often difficult to interpret the model coefficients because of the correlation between the input features. The regression model in this article showed that the collinearity in the data did not seem to have

an effect on the significance and the sign of the regression coefficients for correlated features when the data set is large (>11,000 instances). However, interpreting the actual values of the coefficients is not straightforward because it is unclear to what extent they are influenced by the collinearity. Second, it is difficult to link the regression models to cognitive processes, which receive increasingly more attention in linguistic research.

This motivated our choice for a second approach: a Bayesian Network that exploits the same higher-level features, and of which the graphical structure was based on theoretical reasoning. The network was equally accurate at the classification task as the logistic regression model: 93.2% in 10-fold cross-validation. The major advantages of the Bayesian Network approach are that it enables the modelling of the dependencies between the features explicitly, that it allows introducing hidden nodes that summarise other nodes, and that Bayesian inference can be associated with cognitive processes (Chater et al. 2006). Not only the classification accuracies, but also the classification of the individual constructions was similar to that by logistic regression: the Pearson correlation of the PD-likelihood scores was very high (0.95), and the classes based on these scores differed for only 3.3%: 384 (241 + 143) of the 11,784 instances. For research on alternations in general, a positive aspect of Bayesian Networks is that the number of feature values per node is not limited to two. This means that it is much easier to treat multi-class problems (such as the placement of adverbs in a sentence) than with logistic regression, which allows only pairwise comparisons. One of the risks of Bayesian Networks is that they may introduce circular reasoning. The topology of our networks was based on pre-existing theory and the outcomes of previous experiments. Today, there are no efficient techniques for learning the topology from the data; neither is it easy to determine whether arrows in a network that are mainly responsible for high classification accuracy indeed reflect the underlying cognitive processes. Also, the features on which the networks operate are derived from pre-existing theory. On the other hand, Bayesian Networks can help to falsify existing theories by showing that they cannot explain real (observed) language behaviour.

The accuracy of the memory-based learning approach, making use of lexical items instead of the higher-level features, did not differ significantly from the two approaches making use of higher-level features; in 10-fold cross-validation, the accuracy was 92.5%. However, the classification of the individual cases by the memory-based model differed most from that by the other two approaches, as we saw from the confusion matrices of the classifications and the Pearson correlations of the PD-likelihood scores. The instances that received a different class in memory-based learning than

in the logistic regression model and the Bayesian Network were mostly instances with objects with large variation in words. Apparently, for cases where the possible words in the recipient or theme form a small set (e.g. in cases where it is a pronoun), the classifications are similar to that by the other two models, but for the more unique objects, there are differences. In Section 4.3, we already mentioned that with the current data, it is impossible to say whether these differences are caused by the fact that memory-based learning models do not abstract away from the raw language input (while humans may do so), or by the fact that there is too little data available for the model to be able to classify correctly the less frequent cases (while humans receive many more exemplars). A model of human language acquisition in which language experience is stored and used in new situations, using general cognitive abilities instead of an innate language faculty (as for instance suggested in Daelemans and van den Bosch 2005; Gahl and Yu 2006; Bod 2009), could therefore still be a suitable model for the dative alternation.

Regardless of the type of input features and the type of modelling technique, the largest part of the instances (92.0%) received the same class when training and testing on the same data, most of which (89.2% of all instances) were classified correctly. Seeing that the baseline using only the verb and the length difference already scores an accuracy of 89.6%, and all three approaches used these two features, this high level of agreement is not surprising. Also, we should note that several types of – often somewhat complex – dative constructions (e.g. passive and imperative clauses, clausal objects, etc.) were filtered out in our semi-automatic data collection. The filtering was partly the result of our decision to prevent the influence of other types of syntactic variation (passive versus active voice, declarative versus interrogative mode, the placement of adverbials, etc.). For the other part, they were an artifact of the approach chosen: keeping only those instances for which the higher-level feature values could be established, those that contained a verb in our list of dative verbs, and those that could be detected by the syntactic parser employed. As a result, only the more prototypical instances of the dative alternation are taken into account in this article. It is unclear how including the more complex constructions would have affected the predictive power of the different models considered and the explanatory value of the different higher-level features. Quite likely, including phrasal objects would have complicated the annotation for the higher level features and the feature selection in memory-based learning. Also, it is quite possible that the identity of the verb and the length of the objects are less predictive in the more complex constructions.

Nonetheless, the three full models provide significantly more accurate predictions than the baseline using verb and length difference only. Both the higher-level features and the lexical features may thus play a role in choosing one of the dative constructions. Seeing the small improvement over the baseline, however, it seems that in the data set used, the role of the features is limited and therefore difficult to establish. For now, this means that we cannot be certain that humans make use of abstract semantic properties such as animacy and definiteness when choosing between the two dative constructions. At the same time, it appears that different verbs come with their own preferred constructions, which might give credibility to a theory based on memory-based processes. Also, one may speculate that realising the shortest (and usually given) object first frees memory and processing capacity for articulating the longer (and usually new) one, especially in spontaneous speech.

For the time being, we cannot draw hard and fast conclusions about which modelling technique is best suited to our purposes. Instead of only focussing on the static representations of already produced language (corpus data) as done in this article, research should also be directed at the exploration of models and feature representations that can be more closely linked to cognitive processes in *online* language production. Also, the studies should be extended to other syntactic alternations and other languages, to see how the feature representations and models hold across syntactic constructions and across languages.

Notes

1. This way of residualising means including the pronominality of the recipient and the theme in a linear regression model that predicts length difference. The unexplained variance (the residuals) is then included as a fixed factor in the eventual logistic regression model, replacing the original length difference.
2. Many of the dative verbs are not in the parser lexicon as being dative verbs (and cannot be added as such by users), hence the lower number of verbs (46 instead of 76) in Table 1.
3. To make sure that ungrammatical objects were also classified correctly, we accepted all forms of the pronouns: *I*, *me*, *my*, *mine*, *myself*, *you*, *your*, *yours*, *yourself*, *yourselves*, *we*, *us*, *our*, *ours* and *ourselves*.
4. We use the function *lmer()* (Bates 2005) in *R* (R Development Core Team 2008).
5. The function *lmer()* cannot cope with numerous missing feature value combinations, which is the case with dLenDif78: for 30 of the 78 values, there are $\leq$3 data instances.

6. The accuracy reached when training and testing on the same data is sometimes also referred to as *model fit*, *empirical fit* or *performance ceiling*.
7. See http://dsl.sis.pitt.edu.
8. Again, we thus established the *model fit*.
9. Strictly speaking, this is not a fair train-dev-test split, since we tune on the complete data set (including test data). But since our qualitative evaluation will be based on the models built on all instances, we wanted the variables and parameters of the 10 models in the cross-validation to match those of these models. We believe this decision is defendable because all three approaches have the same benefit.
10. This score was reached with a logistic regression model with verb included as a random effect and length difference (dLenDif5) as the only fixed factor. The type of length difference had no influence on the accuracy reached. Memory-based learning and Bayesian Networks also scored accuracies above 89.0% when provided with only the verb and a form of length difference.
11. We used *collin.fnc()* in the *languageR* package in *R*.
12. We ran TiMBL with *+v db -G0* to obtain these normalised distributions.

References

Arnold, Jennifer, Thomas Wasow, Anthony Losongco, and Ryan Ginstrom.
2000 Heaviness vs. newness: The effects of complexity and information structure on constituent ordering. *Language* 76 (1): 28–55.

Baayen, R. Harald
2008 *Analyzing linguistic data. A practical introduction to statistics using R*. Cambridge: Cambridge University Press.
2011 Corpus linguistics and naive discriminative learning. *Brazilian Journal of Applied Linguistics* 11: 295–328.

Bates, Douglas
2005 Fitting linear mixed models in R. *R News* 5 (1): 27–30.

Behaghel, Otto
1909 Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern [Relationshipsbetween size and ordering of constituents]. *Indogermanische Forschungen* 25: 110–142.

Biber, Douglas
1988 *Variation across speech and writing*. Cambridge: Cambridge University Press.

BNC Consortium
2007 The British National Corpus, version 3 (BNC XML Edition). Available from http://www.natcorp.ox.ac.uk/.

Bod, Rens
2009 From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science* 33 (5): 752–793.

van den Bosch, Antal
2004 Wrapped progressive sampling search for optimizing learning algorithm parameters. In *Proceedings of the 16th Belgian-Dutch Conference on Artificial Intelligence*, Rineke Verbrugge, Niels Taatgen, and Lambert Schomaker (eds.).

Bresnan, Joan, Anna Cueni, Tatiana Nikitina, and R. Harald Baayen
2007 Predicting the dative alternation. In *Cognitive Foundations of Interpretation*, Gerlof Bouma, Irene Kraemer, and Joost Zwarts (eds.), 69–94. Amsterdam, The Netherlands: Royal Netherlands Academy of Science.

Bresnan, Joan, and Marilyn Ford
2010 Predicting syntax: Processing dative constructions in American and Australian varieties of English. *Language* 86 (1): 168–213.

Bresnan, Joan, and Jennifer Hay
2008 Gradient grammar: An effect of animacy on the syntax of *give* in New Zealand and American English. *Lingua* 118 (2): 245–259.

Chater, Nick, Joshua B. Tenenbaum, and Alan Yuille
2006 Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences* 10 (7): 287–291. (Introduction to the special issue on Probabilistic models of cognition.)

Chater, Nick, Mike Oaksford, Ulrike Hahn, and Evan Heit
2010 Bayesian models of cognition. *WIREs Cognitive Science* 1: 811–823.

Daelemans, Walter, and Antal van den Bosch
2005 *Memory-Based Language Processing*. Cambridge, UK: Cambridge University Press.

Daelemans, Walter, Jacub Zavrel, Ko van der Sloot, K., and Antal van den Bosch
2010 TiMBL: Tilburg Memory Based Learner version 6.3 Reference Guide.

Dowman, Mike
2004 Colour terms, syntax and Bayes: Modelling acquisition and evolution. Ph. D. diss., School of Information Technologies, University of Sydney.

Fellbaum, Christiane
1998 *WordNet: An electronic lexical database*. Cambridge, MA, USA: MIT Press.

Gahl, Susanne, and Alan C.L. Yu
2006 Introduction to the special issue on exemplar-based models in linguistics. *The Linguistic Review* 23 (3): 213–216.

Geeraerts, Dirk, Gitte Kristiansen, and Yves Peirsman
2010 *Advances in Cognitive Sociolinguistics*. Berlin, Germany: Walter de Gruyter.

Gries, Stefan Th.
2003 Towards a corpus-based identification of prototypical instances of constructions. *Annual Review of Cognitive Linguistics* 1 (4): 1–27.

Gries, Stefan Th., and Anatol Stefanowitsch
2004 Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9 (1): 97–129.
Grimm, Scott, and Joan Bresnan
2009 Spatiotemporal variation in the dative alternation: a study of four corpora of British and American English. In *Third International Conference Grammar and Corpora.*
Hinrichs, Lars, and Benedikt Szmrecsányi
2007 Recent changes in the function and frequency of Standard English genitive constructions: a multivariate analysis of tagged corpora. *English Language and Linguistics* 11 (3): 437–474.
Jankowski, Bridget
2009 Grammatical and register variation and change: A multicorpora perspective on the English genitive. In *American Association for Corpus Linguistics (AACL 2009).*
Kendall, Tyler, Joan Bresnan, and Gerard van Herk
2011 The dative alternation in African American English: Researching syntactic variation and change across sociolinguistic datasets. *Corpus Linguistics and Linguistic Theory* 7 (2): 229–244.
Kipper, Karin, Hoa Trang Dang, and Martha Palmer
2000 Class-based construction of a verb lexicon. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI-2000),* 691–696.
Koiter, Joost
2006 Visualizing Inference in Bayesian Networks. Master's thesis, Man-machine interaction group, Delft University of Technology.
Levin, Beth
1993 *English verb classes and alternations: A preliminary investigation*. Chicago, IL, USA: The University of Chicago.
McClelland, James L., Matthew M. Botvinick, David C. Noelle, David C. Plaut, Timothy T. Rogers, Mark S. Seidenberg, and Linda B. Smith
2010 Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences* 14: 348 – 356.
Moore, Roger K.
2003 A comparison of the data requirements of automatic speech recognition systems and human listeners. In *Proceedings of EUROSPEECH' 03*, 2582–2584.
Mukherjee, Joybrato, and Sebastian Hoffmann
2006 Describing verb-complementational profiles of New Englishes. *English World-Wide* 27 (2): 147-173.
Pearl, Judea
1988 *Probabilistic reasoning in intelligent systems: Networks of Plausible Inference*. Cambridge, MA, USA: Morgan Kaufmann Publishers Inc.

R Development Core Team
2008 R: A language and environment for statistical computing. Vienna, Austria. Available from http://www.R-project.org.

Shih, Stephanie, and Jason Grafmiller
2011 Weighing in on end weight. In *Annual Meeting of the Linguistic Society of America.*

Szmrecsányi, Benedikt
2010 The English genitive alternation in a cognitive sociolinguistics perspective. In *Advances in Cognitive Sociolinguistics*, Dirk Geeraerts, Gitte Kristiansen, and Yves Peirsman (eds.), 141–166. Berlin, Germany: Walter de Gruyter.

Szmrecsányi, Benedikt, and Lars Hinrichs
2008 Probabilistic determinants of genitive variation in spoken and written English: A multivariate comparison across time, space and genres. In *The dynamics of linguistic variation: Corpus evidence on English past and present*, Terttu Nevalainen, Irma Taavitsainen, Päivi Pahta, and Minna Korhonen (eds.), 291–309. Amsterdam, The Netherlands: John Benjamins.

Tagliamonte, Sali, and Lidia Jarmasz
2008 Variation and change in the English genitive: A sociolinguistic perspective. In *Annual Meeting of the Linguistic Society of America.*

Tapanainen, Pasi, and Timo Järvinen
1997 A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, 64–71.

Theijssen, Daphne
2010 Variable selection in Logistic Regression: The British English dative alternation. In *Interfaces: Explorations in Logic, Language and Computation*, Thomas Icard, and Reinhard Muskens (eds.), 87–101. (Vol. 6211 of Springer Lecture Notes in Artificial Intelligence.)

Theijssen, Daphne, Lou Boves, Hans van Halteren, and Nelleke Oostdijk
2011a Evaluating automatic annotation: Automatically detecting and enriching instances of the dative alternation. *Language Resources and Evaluation*.

Theijssen, Daphne, Hans van Halteren, Lou Boves, and Nelleke Oostdijk.
2011b The more the merrier? How data set size and noisiness affect the accuracy of predicting the dative alternation. In *21st meeting of Computational Linguistics in the Netherlands (CLIN-21)*, University College Ghent, Ghent, Belgium.

Wolk, Chistopher, Joan Bresnan, Anette Rosenbach, and Benedikt Szmrecsányi
2012 Dative and genitive variability in Late Modern English: Exploring cross-constructional variation and change. *Diachronica*. (To appear.)